# **Approach Note: Big Mart Sales Prediction**

The objective of this project was to develop a predictive model to estimate sales of retail products in different Big Mart outlets based on historical sales data. The approach involved systematic steps of data exploration, preprocessing, feature engineering, model experimentation, and evaluation.

### **1. Problem Understanding**

The dataset consisted of item-level sales records with features such as Item\_Weight, Item\_Visibility, Item\_Type, Outlet\_Size, Outlet\_Location\_Type, Outlet\_Type, and others. The target variable was Item\_Outlet\_Sales. The key challenge was handling missing values, categorical variables, and skewed distributions while extracting meaningful signals for sales prediction.

### **2. Data Exploration & Cleaning**

* Performed **exploratory data analysis (EDA)** to understand variable distributions, correlations, and missing data.
* Missing values in Item\_Weight and Outlet\_Size were imputed using mean/mode strategies.
* Normalized inconsistent labels in categorical features (e.g., merging “LF”, “low fat” into “Low Fat”).

### **3. Feature Engineering**

* Encoded categorical variables using **Label Encoding** and **One-Hot Encoding** where applicable.
* Extracted information from identifiers (e.g., Outlet\_Age derived from outlet establishment year).
* Standardized numerical variables where model sensitivity required it.

### **4. Model Building & Experimentation**

Several regression algorithms were tested in an incremental manner:

* **Baseline Models**: Linear Regression
* **Tree-based Models**: Random Forest Regressor to capture non-linear relationships.
* **Advanced Models**: XGBoo.

Hyperparameter tuning was performed (via GridSearchCV) to identify the best configuration for ensemble models.

### **5. Evaluation**

* Used **Root Mean Squared Error (RMSE)** as the primary evaluation metric to measure predictive accuracy.
* Ensemble methods (XGBoost, Random Forest) significantly outperformed linear models, demonstrating their ability to capture complex patterns in sales data.
* Final model selection was based on a balance of accuracy, interpretability, and computational efficiency.

### **6. Key Learnings & Insights**

* Data preprocessing, especially treatment of missing values and categorical encoding, was crucial to performance improvements.
* Feature engineering (outlet age) provided meaningful boosts to prediction power.
* Boosting-based models consistently yielded the best results, making them well-suited for structured retail sales prediction problems.